Mutual Information Bounds via Adjacency Events

Yanjun Han, Or Ordentlich, and Ofer Shayevitz, *Senior Member, IEEE*

***Abstract*— The mutual information between two jointly distributed random variables X and Y is a functional of the joint distribution P_{XY}, which is sometimes difficult to handle or estimate. A coarser description of the statistical behavior of (X, Y) is given by the marginal distributions P_X, P_Y and the *adjacency* relation induced by the joint distribution, where x and y are adjacent if $P(x, y) > 0$. We derive a lower bound on the mutual information in terms of these entities. The bound is obtained by viewing the channel from X to Y as a probability distribution on a set of possible *actions*, where an action determines the output for any possible input, and is independently drawn. We also provide an alternative proof based on convex optimization that yields a generally tighter bound. Finally, we derive an upper bound on the mutual information in terms of adjacency events between the action and the pair (X, Y), where in this case, an action a and a pair (x, y) are adjacent if $y = a(x)$. As an example, we apply our bounds to the binary deletion channel and show that for the special case of an independent identically distributed input distribution and a range of deletion probabilities, our lower and upper bounds both outperform the best known bounds for the mutual information.**

***Index Terms*— Mutual information bounds, functional representation, alternating minimization, deletion channel.**

I. Introduction

THE mutual information $I(X; Y)$ between two jointly distributed random variables X and Y arises as the fundamental limit in many information theoretic problems. When the alphabets $\mathcal{X}$ and $\mathcal{Y}$ are small, the computation of $I(X; Y)$ can be performed directly. This is the typical scenario when considering e.g. the calculation of capacity of memoryless channels, assuming the optimal input distribution is known. In many cases however, the alphabet may become large or even grow unbounded; this is the case e.g. with the capacity of channels with memory that are information stable [1], where the capacity is essentially given by the limit of $I(X^n; Y^n)/n$, for the optimal input X^n. In such cases, it often becomes prohibitively difficult or even virtually impossible to precisely compute the mutual information, hence one must resort to bounding techniques.

In many problems, the marginal distributions of X and Y are simple and the computation of the entropies $H(X)$ and $H(Y)$ is more tractable. In such cases the main obstacle becomes handling the joint distribution and computing the joint (or conditional) entropy. One such prominent example is the binary deletion channel [2] with deletion probability d and an i.i.d. uniform input process. For this setting, the normalized output entropy is easy to derive and approaches $(1 - d)$. However, to evaluate the joint distribution for any given input-output pair, one needs to find the number of different ways the output can be obtained from the input by deleting input bits. This is a difficult combinatorial question, and consequently computing the joint entropy is very challenging. A simpler combinatorial question is to determine whether the output can be obtained from the input by *some* deletion pattern. More generally put, instead of fully characterizing the joint distribution, it is sometimes much easier to characterize its support. Thus, the goal of this work is to provide bounds on the mutual information as a function of the marginals and the joint support. These bounds will be useful when the support is sparse.

In what follows, we assume the alphabets $\mathcal{X}, \mathcal{Y}$ are finite unless otherwise stated. We say that x and y are *adjacent* if $P_{XY}(x, y) > 0$, and we denote this relation by $x \sim y$. We call the event $\mathbb{1}(x \sim y)$ an *adjacency event*. Our first main result is the following.

Theorem 1: For any jointly distributed discrete r.vs (X, Y),

$$I(X, Y) \geq -\mathbb{E}_Y \log \mathbb{E}_X \mathbb{1}(X \sim Y) - \mathbb{E}_X \log \mathbb{E}_Y \frac{\mathbb{1}(X \sim Y)}{\mathbb{E}_X \mathbb{1}(X \sim Y)} \quad (1)$$

Note that by Jensen's inequality both summands are nonnegative, and therefore as a corollary we also get that $I(X, Y) \geq -\mathbb{E}_Y \log \mathbb{E}_X \mathbb{1}(X \sim Y)$. One can find examples where both bounds are tight, e.g., for the mutual information between input and output of the binary erasure channel. It is instructive to note that the weaker bound can be derived directly by the following argument. Draw an i.i.d. codebook with block length n according to P_X, and use it to communicate over a memoryless channel $P_{Y|X}$. Consider the following decoding rule: If the output sequence y^n is P_Y-typical and there is a *unique* codeword x^n such that $x_k \sim y_k$ for all k, output that codeword. Otherwise, declare an error. Clearly, $\Pr(X_k \sim y_k) = E_X \mathbb{1}(X \sim y_k)$ and thus, assuming that y^n is typical, the probability (averaged over random codebooks) that a specific codeword will satisfy the decoding rule is $\approx \prod_{y \in \mathcal{Y}} (\mathbb{E}_X \mathbb{1}(X \sim y))^{nP(y)}$. Therefore, by the union bound, any rate below $-\mathbb{E}_Y \log \mathbb{E}_X \mathbb{1}(X \sim Y)$ can be attained by this strategy with vanishing error probability, and this in turn cannot be larger than the mutual information. A bound

Manuscript received July 22, 2016; revised August 23, 2016; accepted September 4, 2016. Date of publication September 14, 2016; date of current version October 18, 2016. Y. Han was supported by the NSF Center for Science of Information under Grant CCF-0939370. O. Ordentlich was supported by the MIT-Technion postdoctoral fellowship. O. Shayevitz was supported in part by an ERC under Grant 639573, and an ISF under Grant 1367/14. This paper was presented at the 2014 International Symposium on Information Theory.

Y. Han is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA (email: yjhan@stanford.edu).

O. Ordentlich is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA (email: ordent@mit.edu).

O. Shayevitz is with the Department of EESystems, Tel Aviv University, Tel Aviv 6997801, Israel (email: ofersha@eng.tau.ac.il).

Communicated by H. H. Permuter, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2016.2609390

of this type was implicitly used in [3] and [4]. Our main contribution is therefore the second term in (1). As we shall see in Section V, this additional term can be significant.

Let us briefly provide the main ideas behind our approach. A channel is traditionally defined via a conditional probability distribution $P_{Y|X}$ of the output given the input. Alternatively, a channel can also be (nonuniquely) defined as a random mapping $Y = A(X)$ from an input alphabet to an output alphabet, where the actual mapping applied to the input, namely the channel *action* A, is drawn according to some probability distribution P_A over the set of all possible actions, *independently* of the input (see the functional representation lemma in [5, Appendix B]). Following this paradigm, the mutual information for a given input distribution P_X can be written as

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - (H(Y,A|X) - H(A|X,Y)) \\ &= H(Y) - H(A|X) - H(Y|A,X) + H(A|X,Y) \\ &= H(Y) - H(A) + H(A|X,Y) \end{aligned} \tag{2}$$

where (2) follows since the action A is statistically independent of the input X, and $Y = A(X)$. This holds for any eligible choice of action A. A natural quantity to consider is therefore the *intrinsic uncertainty* $H(A|X,Y)$ associated with A, that captures the amount of information regarding the channel action revealed by observing its input and output. Note that for any eligible choice of A, we have that $I(A;X,Y) = H(A) - H(A|X,Y) = H(Y|X)$ is fixed, but the entropy of the action $H(A)$ and the intrinsic uncertainty associated with the action can vary.

As an example, consider the binary symmetric channel (BSC) with crossover probability $0 < p < \frac{1}{2}$. A natural choice for the action A is drawing a r.v. $Z \sim \text{Bern}(p)$ and setting $A(X) = X \oplus Z$. In this case, the entropy of the action is $H(A) = h(p)$, where $h(\cdot)$ is the binary entropy function, and the intrinsic uncertainty $H(A|X,Y) = 0$, since viewing X and Y completely reveals the action (the noise Z). Another possible choice for the action A is drawing a ternary r.v. U with $\Pr(U = 0) = \Pr(U = 1) = p$, and $\Pr(U = 2) = 1 - 2p$, and setting

$$A(X) = U \cdot \mathbb{1}(U \neq 2) + X \cdot \mathbb{1}(U = 2)$$

In this case, the entropy of the action is $H(A) = h(2p) + 2p$, and the intrinsic uncertainty is $H(A|X,Y) = (1-p) \cdot h\left(\frac{p}{1-p}\right) > 0$, since if $X = Y$ there remains some uncertainty regarding the action. Indeed, it can be directly verified that the identity $h(p) = h(2p) + 2p - (1-p) \cdot h\left(\frac{p}{1-p}\right)$ holds.

Following the above, in Section II we derive a lower bound on the intrinsic uncertainty for any given choice of the action A. This bound is based on an application of the Donsker-Varadhan variational principle. This will immediately translate into lower bounds on the mutual information. Our general statement, given in Theorem 3, is a family of bounds that depend on the particular choice of the action. While these bounds may be generally difficult to evaluate, we show in Section III that for any channel $P_{Y|X}$ there always exists a specific choice of action, such that the associated bound depends only on the marginals and the joint support. This yields Theorem 1.

The proof of Theorem 1 as delineated above in based on information theoretic arguments. Alternatively, the theorem can also be proved more directly using convex optimization techniques. In fact, this alternative approach does not only recover Theorem 1, but can also yield an increasing sequence of bounds that converges to the best possible lower bound on the mutual information in terms of the marginals P_X, P_Y and the support of P_{XY}. Furthermore, while the information theoretic proof applies only to finite alphabets, the convex optimization approach can also handle countably infinite alphabets. This result appears in Theorem 4, Section IV. We note that the improved bounds obtained by this procedure seem quite difficult to evaluate in general.

Interestingly, while actions were introduced in order to lower bound the mutual information, our results can be trivially leveraged to obtain upper bounds as well.

Theorem 2: Let $(X,Y) \sim P_X \times P_{Y|X}$ be jointly distributed discrete r.vs. Let A be any action consistent with $P_{Y|X}$, i.e., such that $Y = A(X)$. Then

$$\begin{aligned} I(X;Y) \leq H(Y) &+ \mathbb{E}_A \log \mathbb{E}_{X,Y} \mathbb{1}(A \sim (X,Y)) \\ &+ \mathbb{E}_{X,Y} \log \mathbb{E}_A \frac{\mathbb{1}(A \sim (X,Y))}{\mathbb{E}_{X,Y} \mathbb{1}(A \sim (X,Y))} \end{aligned} \tag{3}$$

Proof: By (2) we have that $I(X;Y) = H(Y) - I(A;X,Y)$. The proof follows by applying Theorem 1 to $I(A;X,Y)$. ■

Note that $\mathbb{1}(A \sim (X,Y))$ is an indicator on the event where $A(X) = Y$. In (3), the expectations are taken with respect to $(X,Y) \sim P_{XY}$ and $A \sim P_A$ independent of (X,Y). Observe also that both the second and third terms in (3) are non-positive, hence the bound holds even if one of them is removed.

Lastly, in Section V we illustrate the applicability of our bounds in several specific examples. In particular, we provide simple examples showing that our bounds are sometimes tight, and demonstrating that the second term in (1) can be significant. We then consider the binary deletion channel for which the value of the mutual information is currently unknown for any nontrivial input process. For an i.i.d. uniform input, we evaluate our lower and upper bounds, and show that they both outperform the best known bounds on the mutual information. Finally, we draw a relation between the upper bound from Theorem 2 and a recent conjecture of Courtade and Kumar [6]. As all examples we consider in this paper involve binary channels, unless stated otherwise, all logarithms are taken to base 2.

II. A Family of Bounds via Actions

In this section we define a channel by its action on its input, and develop general lower bounds on the mutual information between the input and output in terms of the channel action, by bounding the associated intrinsic uncertainty defined below.

A. Channels via Actions

Let $\mathcal{X}, \mathcal{Y}$ be discrete alphabets. Any channel $P_{Y|X}$ from $\mathcal{X}$ to $\mathcal{Y}$ can be (nonuniquely) defined by a probability distribution P_A on a set $\mathcal{A}$ of mappings from $\mathcal{X} \mapsto \mathcal{Y}$, to which we refer to below as *actions*. Each action $a(\cdot) \in \mathcal{A}$ is defined for all possible inputs, and the channel action is chosen independently of the input, yielding the output $Y = A(X) \in \mathcal{Y}$.

For any eligible choice of action A, the *intrinsic uncertainty* of the channel with respect to the input distribution P_X is defined to be $H(A|X, Y)$. Note that while the intrinsic uncertainty may depend on the choice of A, the difference $H(A) - H(A|X, Y)$, which was shown in Section I to be equal to $H(Y|X)$, does not; we therefore have the freedom to choose the action distribution that is most convenient to work with.

Example 1 (Generic Action Set): For any channel $P_{Y|X}$ we can always generate the action according to the following procedure. Let $\mathcal{A}$ consist of all $|\mathcal{Y}|^{|\mathcal{X}|}$ functions from $\mathcal{X} \mapsto \mathcal{Y}$, and for any $a \in \mathcal{A}$ set $P_A(a) = \prod_{x\in\mathcal{X}} P_{Y|X}(a(x)|x)$. Drawing A according to P_A, statistically independent of X, and setting $Y = A(X)$, is equivalent to drawing in advance a sequence of statistically independent r.vs $\{Y_x\}_{x\in\mathcal{X}}$, where $Y_x \sim P_{Y|X}(\cdot|x)$, and then when X is revealed, outputting only the corresponding Y_X. Thus, the above $\mathcal{A}$ and P_A are consistent with $P_{Y|X}$, i.e., they describe the channel $P_{Y|X}$.

We further note that it is always possible to construct an action set with less than $|\mathcal{X}| \cdot |\mathcal{Y}|$ actions, see the functional representation lemma in [5, Appendix B]. Moreover, in many cases there exist "natural" choices of an action that describes the channel. In Section I we described such choices for the BSC. Below we provide a few more examples.

Example 2 (Z Channel): The (symmetric) Z channel has a binary input X and binary output Y, such that $\Pr(Y = 0|X = 0) = 1$ and $\Pr(Y = 0|X = 1) = \Pr(Y = 1|X = 1) = \frac{1}{2}$. A natural choice for the action A is taking the action set $\mathcal{A}$ to consist of the two actions $a_1(x) = x$ and $a_2(x) = 0$ with probability assignment $p(a_1) = p(a_2) = \frac{1}{2}$.

Example 3 (Deletion Channel): In a deletion channel, each transmitted symbol is either deleted or received uncorrupted. Assuming the input to the channel is an n-dimensional vector $\mathbf{X}$, the set $\mathcal{A}$ includes 2^n actions, each corresponding to a different subset of the input indices $[1 : n]$ marked for deletion. In an i.i.d. deletion model symbols are independently deleted with probability d. Therefore the probability of an action a that deletes exactly w bits is $P(a) = d^w(1-d)^{n-w}$. Different actions applied to the same input may result in the same output. For example, if $\mathbf{x} = 01100$ we may get the output $\mathbf{y} = 110$ if either the first and fourth symbols or the first and fifth symbols were deleted. Therefore, the intrinsic uncertainty $H(A|\mathbf{X}, \mathbf{Y})$ is generally positive.

Example 4 (Trapdoor Channel): The trapdoor channel is a simple finite-state binary channel, defined as follows. Balls labeled "0" or "1" are used to communicate through the channel. The channel starts with a ball already in it, referred to as the initial state. On each channel use, a ball is inserted into the channel by the transmitter, and one of the two balls in the channel is emitted with equal probability. The ball that is not emitted remains inside for the next channel use. In this model, the channel's action consists of choosing the initial state and deciding for each channel use whether to emit the ball that was already inside the channel or the ball that has just entered. Since an input $\mathbf{x}$ can be mapped to an output $\mathbf{y}$ via multiple actions, the intrinsic uncertainty is generally positive.

B. Bounds

Our main tool in lower bounding the intrinsic uncertainty is the variational formula of Donsker and Varadhan (See [7, Ch. 1.4]). We write $D(P\|Q)$ for the relative entropy between the distributions P, Q, and $Q \ll P$ if $P(x) = 0$ implies $Q(x) = 0$.

Lemma 1 (Donsker-Varadhan): For any distribution P and any nonnegative function $f(x)$ for which $\mathbb{E}_P \log f(X)$ is finite,

$$\mathbb{E}_P \log f(X) = \min_{Q\ll P} \log \mathbb{E}_Q f(X) + D(P\|Q), \tag{4}$$

and the minimum is uniquely attained by

$$Q^*(x) = \frac{P(x)/f(x)}{\mathbb{E}_P(1/f(X))}, \tag{5}$$

where by convention we set $1/f(x) = 0$ if $f(x) = 0$.

For completeness, we bring the proof of this lemma.

Proof: Let $Q^*(x)$ be as above. For any distribution Q we have

$$\begin{aligned}
&D(P\|Q) + \log \mathbb{E}_Q f(X)\\
&\quad= \mathbb{E}_P \log \frac{P}{Q} + \log \mathbb{E}_Q f(X)\\
&\quad= \mathbb{E}_P \log \frac{Q^*}{Q} + \mathbb{E}_P \log \frac{P}{Q^*} + \log \mathbb{E}_Q f(X)\\
&\quad= \sum_x P(x) \log \frac{P(x)}{Q(x)f(x)\mathbb{E}_P(1/f(X))}\\
&\qquad+ \mathbb{E}_P \log \frac{P(X)f(X)\mathbb{E}_P(1/f(X))}{P(X)} + \log \mathbb{E}_Q f(X)\\
&\quad\overset{(a)}{\geq} \left(\sum_x P(x)\right) \log \frac{\sum_x P(x)}{\sum_x Q(x)f(x)\mathbb{E}_P(1/f(X))}\\
&\qquad+ \mathbb{E}_P \log f(X) + \log \mathbb{E}_P \frac{1}{f(X)} + \log \mathbb{E}_Q f(X)\\
&\quad= \mathbb{E}_P \log f(X)
\end{aligned}$$

where (a) follows from the log-sum inequality [8, Ch. 2.7] which is tight if and only if $Q(x) = Q^*(x)$. ■

We would like to obtain an alternative expression for

$$H(A|X, Y) = \mathbb{E} \log \frac{1}{P(A|X, Y)}, \tag{6}$$

where the expectation is taken with respect to the joint distribution

$$\begin{aligned}
P(x, y, a) &= P(x)P(a|x)P(y|x, a)\\
&= P(x)P(a)\mathbb{1}(y = a(x)),
\end{aligned}$$

and $\mathbb{1}(B)$ is an indicator function for the event B. For brevity, we sometimes refer to this distribution as P.

Define the distribution

$$Q(x, y, a) \triangleq \frac{P(x, y, a)P(a|x, y)}{\mathbb{E}_P P(A|X, Y)}, \tag{7}$$

which we sometimes refer to as Q. Using the Donsker-Varadhan variational principle with $f(x,y,a) = 1/P(a|x,y)$, the expectation from (6) can be written as

$$\begin{aligned}\mathbb{E}\log\frac{1}{P(A|X,Y)} &= \log\mathbb{E}_Q\frac{1}{P(A|X,Y)} + D(P\|Q)\\ &= \log\mathbb{E}_Q\frac{1}{P(A|X,Y)} + D\left(P_Y\|Q_Y\right)\\ &\quad + D\left(P_{X,A|Y}\|Q_{X,A|Y} \mid P_Y\right), \end{aligned} \tag{8}$$

where (8) follows from the chain rule of relative entropy. The marginal distribution $Q(y)$ is given by

$$\begin{aligned}Q(y) &= \sum_{x,a} Q(x,y,a)\\ &= \frac{1}{\mathbb{E}_P P(A|X,Y)}\sum_{x,a} P(x)P(a)\mathbb{1}(y=a(x))P(a|x,y)\\ &= \frac{\mathbb{E}_{X,A}P(A|X,y)}{\mathbb{E}_P P(A|X,Y)}, \end{aligned} \tag{9}$$

where in (9) we have used the fact that $P(a|x,y) = 0$ whenever $y \neq a(x)$. Thus,

$$\begin{aligned}D\left(P_Y\|Q_Y\right) &= \mathbb{E}_Y\log\left(\frac{P(Y)\mathbb{E}_P P(A|X,Y)}{\mathbb{E}_{X,A}P(A|X,Y)}\right)\\ &= -H(Y) + \log\mathbb{E}_P P(A|X,Y)\\ &\quad - \mathbb{E}_Y\log\mathbb{E}_{X,A}P(A|X,Y). \end{aligned} \tag{10}$$

In addition,

$$\begin{aligned}\log\mathbb{E}_Q\frac{1}{P(A|X,Y)} &= \log\sum_{x,y,a}\frac{Q(x,y,a)}{P(a|x,y)}\\ &= -\log\mathbb{E}_P P(A|X,Y). \end{aligned} \tag{11}$$

Substituting (10) and (11) into (8) yields

$$\begin{aligned}H(A|X,Y) &= -H(Y) - \mathbb{E}_Y\log\mathbb{E}_{X,A}P(A|X,Y)\\ &\quad + D\left(P_{X,A|Y}\|Q_{X,A|Y} \mid P_Y\right). \end{aligned} \tag{12}$$

We are left with the task of evaluating the conditional relative entropy in (12). The conditional distributions that participate in this term are given by

$$P(x,a|y) = P(x)P(a)\frac{\mathbb{1}(y=a(x))}{E_{X,A}\mathbb{1}(y=A(X))} \tag{13}$$

$$Q(x,a|y) = P(x)P(a)\frac{P(a|x,y)}{E_{X,A}P(A|X,y)} \tag{14}$$

and therefore

$$\begin{aligned}&D\left(P_{X,A|Y}\|Q_{X,A|Y} \mid P_Y\right)\\ &\quad = \mathbb{E}_P\log\left(\frac{\mathbb{1}(Y=A(X))}{E_{X,A}\mathbb{1}(Y=A(X))}\cdot\frac{E_{X,A}P(A|X,Y)}{P(A|X,Y)}\right). \end{aligned} \tag{15}$$

Unfortunately, an exact computation of (15) involves the computation of $\mathbb{E}_P\log(1/P(A|X,Y))$, which is the exact technical difficulty we are trying to avoid. Instead, we lower bound (15) using the convexity of relative entropy, i.e.,

$$D\left(P_{X,A|Y}\|Q_{X,A|Y} \mid P_Y\right) \geq D\left(P_{X,A}\|\tilde{Q}_{X,A}\right), \tag{16}$$

where

$$\begin{aligned}\tilde{Q}(x,a) &= \sum_y P(y)Q(x,a|y)\\ &= P(x,a)\mathbb{E}_Y\frac{P(a|x,Y)}{E_{X,A}P(A|X,Y)}. \end{aligned} \tag{17}$$

Note that other properties of relative entropy, such as the data-processing inequality or Pinsker's inequality, could potentially be useful for bounding (15). Combining (16) and (17) gives,

$$D\left(P_{X,A|Y}\|Q_{X,A|Y} \mid P_Y\right) \geq -\mathbb{E}_{X,A}\log\mathbb{E}_Y\frac{P(A|X,Y)}{E_{X,A}P(A|X,Y)}. \tag{18}$$

Substituting (18) into (12) and using (2) yields the following.

Theorem 3: Let $(X,Y) \sim P_X \times P_{Y|X}$ be jointly distributed discrete r.vs. Let A be any action consistent with $P_{Y|X}$, i.e., such that $Y = A(X)$. Then

$$\begin{aligned}I(X;Y) &\geq -H(A) - \mathbb{E}_Y\log\mathbb{E}_{X,A}P(A|X,Y)\\ &\quad - \mathbb{E}_{X,A}\log\mathbb{E}_Y\frac{P(A|X,Y)}{E_{X,A}P(A|X,Y)}. \end{aligned} \tag{19}$$

III. A Bound via Adjacency Events

An action A is called *uniform* if all actions in its support $\mathcal{A}$ are equiprobable, i.e.,

$$P(a) = \begin{cases}\frac{1}{|\mathcal{A}|} & a \in \mathcal{A}\\ 0 & a \notin \mathcal{A}.\end{cases}$$

At this point, we restrict our attention to this class of actions, for which the bound in Theorem 3 takes a particularly simpler form that depends only on the marginal distributions of X and Y and their joint support. We then show that any channel can be essentially characterized by a uniform action, which in turn proves Theorem 1.

For any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ let

$$\mathcal{A}(x,y) \triangleq \{a \ : \ a(x) = y\} \tag{20}$$

be the set of all possible actions in $\mathcal{A}$ that map the input x to the output y. Denote the cardinality of this set by $N(x,y) \triangleq |\mathcal{A}(x,y)|$.

Proposition 1: If A is a uniform action, then A conditioned on X and Y is uniformly distributed over the set $\mathcal{A}(X,Y)$.[1]

Proof:

$$\begin{aligned}P(a|x,y) &= \frac{P(x,y|a)P(a)}{P(x,y)}\\ &= \frac{P(y|x,a)P(a)}{P(y|x)}\\ &= \frac{\mathbb{1}(y=a(x))P(a)}{\sum_{a\in\mathcal{A}(x,y)}P(a)}\\ &\overset{(a)}{=} \frac{\frac{1}{|\mathcal{A}|}\mathbb{1}\left(a\in\mathcal{A}(x,y)\right)}{\frac{1}{|\mathcal{A}|}N(x,y)}\\ &= \frac{\mathbb{1}\left(a\in\mathcal{A}(x,y)\right)}{N(x,y)}, \end{aligned} \tag{21}$$

[1] Note that the converse is not generally true. As a counterexample, consider the BSC with the action $A(X) = X \oplus Z$ where $Z \sim \text{Bern}(p)$.

where (a) follows from $\mathbb{1}(y = a(x)) = \mathbb{1}(a \in \mathcal{A}(\mathbf{x}, \mathbf{y}))$ and since $P(a) = \frac{1}{|\mathcal{A}|}$ for all $a \in \mathcal{A}$. ■

Lemma 2: Suppose $P_{Y|X}$ can be represented by a uniform action A. Then, for any input distribution P_X

$$-\mathbb{E}_Y \log \mathbb{E}_{X,A} P(A|X,Y) = H(A) - \mathbb{E}_Y \log \mathbb{E}_X \mathbb{1}(X \sim Y). \tag{22}$$

Proof: Using Proposition 1,

$$\begin{aligned}\mathbb{E}_{X,A} P(A|X,y) &= \sum_x P(x) \sum_a P(a) \frac{\mathbb{1}\,(a \in \mathcal{A}(x,y))}{N(x,y)} \\ &= \frac{1}{|\mathcal{A}|} \sum_x P(x) \frac{N(x,y)}{N(x,y)} \mathbb{1}(x \sim y) \\ &= \frac{1}{|\mathcal{A}|} \mathbb{E}_X \mathbb{1}(X \sim y).\end{aligned} \tag{23}$$

Thus,

$$\begin{aligned}-\mathbb{E}_Y \log \mathbb{E}_{X,A} P(A|X,Y) &= -\mathbb{E}_Y \log \frac{1}{|\mathcal{A}|} \\ &\quad - \mathbb{E}_Y \log \mathbb{E}_X \mathbb{1}(X \sim Y) \\ &= \log|\mathcal{A}| - \mathbb{E}_Y \log \mathbb{E}_X \mathbb{1}(X \sim Y).\end{aligned}$$

The lemma follows since $H(A) = \log|\mathcal{A}|$ by the assumption that A is a uniform action. ■

The next lemma lower bounds the last term in (19) for channels with a uniform action A.

Lemma 3: Suppose $P_{Y|X}$ can be represented by a uniform action A. Then, for any input distribution P_X

$$\begin{aligned}&-\mathbb{E}_{X,A} \log \mathbb{E}_Y \frac{P(A|X,Y)}{E_{X,A} P(A|X,Y)} \\ &\quad \geq -\mathbb{E}_X \log \mathbb{E}_Y \frac{\mathbb{1}(X \sim Y)}{\mathbb{E}_X \mathbb{1}(X \sim Y)} \geq 0\end{aligned} \tag{24}$$

Proof: By virtue of Jensen's inequality,

$$\begin{aligned}&-\mathbb{E}_{X,A} \log \mathbb{E}_Y \frac{P(A|X,Y)}{E_{X,A} P(A|X,Y)} \\ &\quad \geq -\mathbb{E}_X \log \mathbb{E}_Y \frac{\mathbb{E}_A P(A|X,Y)}{E_{X,A} P(A|X,Y)}.\end{aligned}$$

Using (21) and (23), we have

$$\begin{aligned}\frac{\mathbb{E}_A P(A|x,y)}{E_{X,A} P(A|X,y)} &= \frac{\sum_a P(a) \frac{\mathbb{1}(a \in \mathcal{A}(x,y))}{N(x,y)}}{\frac{1}{|\mathcal{A}|} \mathbb{E}_X \mathbb{1}(X \sim y)} \\ &= \frac{\mathbb{1}(x \sim y)}{\mathbb{E}_X \mathbb{1}(X \sim y)},\end{aligned}$$

establishing the first inequality in (24). The second inequality follows by applying Jensen's inequality again, this time w.r.t. $\mathbb{E}_X$. ■

Combining Theorem 3, Lemma 2, and Lemma 3, establishes the following.

Lemma 4: Suppose $P_{Y|X}$ can be represented by a uniform action A. Then, for any input distribution P_X

$$\begin{aligned}I(X;Y) &\geq -\mathbb{E}_Y \log \mathbb{E}_X \mathbb{1}(X \sim Y) \\ &\quad - \mathbb{E}_X \log \mathbb{E}_Y \frac{\mathbb{1}(X \sim Y)}{\mathbb{E}_X \mathbb{1}(X \sim Y)}.\end{aligned} \tag{25}$$

To establish our main result for any channel and input distribution, we first show the following.

Lemma 5: Let $P_{Y|X}$ be a channel with the property that $P(y|x)$ is rational for all x and y. Then there exists a uniform action for $P_{Y|X}$.

Proof: For any channel $P_{Y|X}$ with rational probabilities there exists some action set $\mathcal{A} = \{a_1, \cdots, a_{|\mathcal{A}|}\}$ and a corresponding probability distribution P_A consistent with it such that all probabilities $P_A(a_i)$, $i = 1, \ldots, |\mathcal{A}|$, are positive rational numbers. For example, the construction from Example 1 yields rational probabilities $P_A(a_i)$, $i = 1, \ldots, |\mathcal{A}|$. We construct a new action $\bar{A}$ by duplicating each action a_i to M_i identical actions, and assigning the probability $P_A(a_i)/M_i$ to each of them. Clearly, the new action is also consistent with $P_{Y|X}$ for any choice of the natural numbers $M_1, \ldots, M_{|\mathcal{A}|}$. By our assumption that all original action probabilities are positive rational numbers, we can always find a choice of $M_1, \ldots, M_{|\mathcal{A}|}$ such that all new action probabilities are equal. For such a choice the action $\bar{A}$ will be uniform. ■

Using Lemma 4 and Lemma 5, we can now prove our main result.

Proof of Theorem 1: Any channel $P_{Y|X}$ can be approximated arbitrarily well by a conditional distribution $\tilde{P}_{\tilde{Y}|X}$ with the same support whose entries are all rational, in the sense that $\max_{x,y} |P_{Y|X}(y|x) - \tilde{P}_{Y|X}(y|x)|$ can be made arbitrarily small. This means that both $P_X \times \tilde{P}_{Y|X}$ and the corresponding marginal $\tilde{P}_Y$ are arbitrarily close to P_{XY} and P_Y respectively. Since the mutual information $I(X;Y)$ is continuous with respect to P_{XY}, the mutual information $I(X;\tilde{Y})$ between X and the output of the "rational" channel $\tilde{P}_{Y|X}$ can be made arbitrarily close to $I(X;Y)$. By Lemma 5, there exists a uniform action for $\tilde{P}_{\tilde{Y}|X}$, and consequently by Lemma 4 its mutual information is lower bounded by (25). By continuity, $I(X;Y)$ is also lower bounded by (25). ■

IV. A Convex–Optimization Based Bound

In the previous section we have proved a lower bound on $I(X;Y)$ that depends only on the marginal distributions P_X, P_Y and the support of the joint distribution, namely, the function $\mathbb{1}(x \sim y)$. Our proof relied on information theoretic arguments. In this section we will take a more direct approach to the problem, and derive bounds on $I(X;Y)$ in terms of the same quantities, using convex optimization. More specifically, to arrive at a lower bound we minimize $I(X;Y)$ w.r.t. P_{XY} subject to the constraints that the marginal distributions are P_X, P_Y, and that $P_{XY}(x,y) = 0$ whenever $\mathbb{1}(x \sim y) = 0$. Throughout this section we assume all logarithms are in the natural basis, while the result of Theorem 4 remains valid as long as the same logarithmic basis is applied to $I(X;Y)$.

We consider the following problem:

$$\begin{aligned}&\text{minimize} \quad I(X;Y) \\ &\text{subject to:} \\ &\qquad \sum_{x:x\sim y} P_{XY}(x,y) = P_Y(y) \qquad \forall y \in \mathcal{Y} \\ &\qquad \sum_{y:y\sim x} P_{XY}(x,y) = P_X(x) \qquad \forall x \in \mathcal{X} \\ &\qquad P_{XY}(x,y) \geq 0 \quad \text{if } x \sim y, \\ &\qquad P_{XY}(x,y) = 0 \text{ if } x \nsim y.\end{aligned}$$

Note that the constraints above imply $\sum_{x,y} P_{XY}(x,y) = 1$. This is equivalent to

$$\begin{aligned}
&\text{minimize } \sum_{x\sim y} P_{XY}(x,y)\log\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}\\
&\text{subject to:}\\
&\qquad \sum_{x:x\sim y} P_{XY}(x,y) = P_Y(y) \qquad \forall y\in\mathcal{Y}\\
&\qquad \sum_{y:y\sim x} P_{XY}(x,y) = P_X(x) \qquad \forall x\in\mathcal{X}\\
&\qquad P_{X,Y}(x,y)\geq 0 \qquad \forall (x,y)\in\mathcal{X}\times\mathcal{Y},\ x\sim y
\end{aligned}$$

This objective function is convex in $P_{XY}(x,y)$, and the constraints are linear, so the optimization solution can be obtained by the solution to the dual problem given by

$$\begin{aligned}
L &= \inf_{P_{XY}(x,y)} \sup_{\lambda_x,\mu_y\in\mathbb{R},\tau_{xy}\geq 0} \sum_{x\sim y} P_{XY}(x,y)\log\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}\\
&\quad - \sum_x \lambda_x\left(\sum_{y:y\sim x} P_{XY}(x,y) - P_X(x)\right)\\
&\quad - \sum_y \mu_y\left(\sum_{x:x\sim y} P_{XY}(x,y) - P_Y(y)\right)\\
&\quad - \sum_{x\sim y} \tau_{xy}P_{XY}(x,y)\\
&= \sup_{\lambda_x,\mu_y\in\mathbb{R},\tau_{xy}\geq 0} \inf_{P_{XY}(x,y)} \sum_x \lambda_x P_X(x) + \sum_y \mu_y P_Y(y)\\
&\quad + \sum_{x\sim y} P_{XY}(x,y)\left(\log\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} - \lambda_x - \mu_y - \tau_{xy}\right) \qquad (26)\\
&= \sup_{\lambda_x,\mu_y\in\mathbb{R},\tau_{xy}\geq 0} \inf_{P_{XY}(x,y)} \sum_x \lambda_x P_X(x) + \sum_y \mu_y P_Y(y)\\
&\quad + \sum_{x\sim y} P_{XY}(x,y)\left(\log P_{XY}(x,y) - a_{xy}\right)
\end{aligned}$$

where (26) follows from the minimax theorem and

$$a_{xy} \triangleq \log(P_X(x)P_Y(y)) + \lambda_x + \mu_y + \tau_{xy}.$$

The function $f(x) = x\log x - ax$ is minimized at $x^* = e^{a-1}$ and its minimal value is $f(x^*) = -e^{a-1}$. Using this, we get that

$$\begin{aligned}
L &= \sup_{\lambda_x,\mu_y\in\mathbb{R},\tau_{xy}\geq 0} \sum_{x\sim y} -e^{\log(P_X(x)P_Y(y))+\lambda_x+\mu_y+\tau_{xy}-1}\\
&\quad + \sum_x \lambda_x P_X(x) + \sum_y \mu_y P_Y(y)\\
&= \sup_{\lambda_x,\mu_y\in\mathbb{R},\tau_{xy}\geq 0} \sum_{x\sim y} -P_X(x)P_Y(y)e^{\lambda_x+\mu_y+\tau_{xy}-1}\\
&\quad + \sum_x \lambda_x P_X(x) + \sum_y \mu_y P_Y(y)
\end{aligned}$$

Clearly, the maximizing τ_{xy} is $\tau_{xy} = 0$ which gives

$$\begin{aligned}
L &= \sup_{\lambda_x,\mu_y\in\mathbb{R}} \sum_{x\sim y} -P_X(x)P_Y(y)e^{\lambda_x+\mu_y-1}\\
&\quad + \sum_x \lambda_x P_X(x) + \sum_y \mu_y P_Y(y)\\
&= \sup_{\lambda_x,\mu_y\in\mathbb{R}} \sum_{x\sim y} -P_X(x)P_Y(y)e^{\lambda_x+\mu_y}\\
&\quad + \sum_x \lambda_x P_X(x) + \sum_y \mu_y P_Y(y) + 1
\end{aligned}$$

where in the last step we replaced μ_y with $\mu_y - 1$ (with some abuse of notation). Let $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ be the vectors holding $\{\lambda_x\}_{x\in\mathcal{X}}$ and $\{\mu_y\}_{y\in\mathcal{Y}}$, respectively, and

$$\begin{aligned}
G(\boldsymbol{\lambda},\boldsymbol{\mu}) &\triangleq \sum_{x\sim y} -P_X(x)P_Y(y)e^{\lambda_x+\mu_y}\\
&\quad + \sum_x \lambda_x P_X(x) + \sum_y \mu_y P_Y(y) + 1,
\end{aligned}$$

such that

$$L = \sup_{\boldsymbol{\lambda}\in\mathbb{R}^{|\mathcal{X}|},\boldsymbol{\mu}\in\mathbb{R}^{|\mathcal{Y}|}} G(\boldsymbol{\lambda},\boldsymbol{\mu}).$$

We will use the alternating minimization approach to minimize $-G(\boldsymbol{\lambda},\boldsymbol{\mu})$ (which is equivalent to maximizing $G(\boldsymbol{\lambda},\boldsymbol{\mu})$) over $\mathbb{R}^{|\mathcal{X}|}\times\mathbb{R}^{|\mathcal{Y}|}$). This approach is described as follows: for arbitrary initialization of $\boldsymbol{\lambda}^{(0)}$, we use an iterative algorithm to successively minimize the target function. In k-th iteration, we first hold $\boldsymbol{\lambda}^{(k-1)}$ fixed and minimize the target function over $\boldsymbol{\mu}$ to obtain $\boldsymbol{\mu}^{(k-1)}$, and then hold $\boldsymbol{\mu}^{(k-1)}$ fixed and minimize the the target function over $\boldsymbol{\lambda}$ to obtain $\boldsymbol{\lambda}^{(k)}$. In mathematical forms, for $k\geq 0$, we have

$$\begin{aligned}
\boldsymbol{\mu}^{(k)} &\in \operatorname*{argmin}_{\boldsymbol{\mu}} -G(\boldsymbol{\lambda}^{(k)},\boldsymbol{\mu}),\\
\boldsymbol{\lambda}^{(k+1)} &\in \operatorname*{argmin}_{\boldsymbol{\lambda}} -G(\boldsymbol{\lambda},\boldsymbol{\mu}^{(k)}).
\end{aligned}$$

The alternating minimization approach is widely used in optimization where separate optimization over different parameter subsets is much easier than the joint optimization, e.g., in the expectation–minimization (EM) algorithm [9] to find the maximum likelihood estimator, in the Blahut–Arimoto algorithm [10], [11] to maximize the mutual information between channel input and output, in minimizing the Kullback–Leibler divergence between two convex sets of finite measures [12], to name a few. One remarkable property of this approach is that, by definition we have

$$\begin{aligned}
G(\boldsymbol{\lambda}^{(0)},\boldsymbol{\mu}^{(0)}) &\leq G(\boldsymbol{\lambda}^{(1)},\boldsymbol{\mu}^{(0)}) \leq G(\boldsymbol{\lambda}^{(1)},\boldsymbol{\mu}^{(1)})\\
&\leq G(\boldsymbol{\lambda}^{(2)},\boldsymbol{\mu}^{(1)}) \leq \cdots \leq L \qquad (27)
\end{aligned}$$

i.e., the value sequence obtained by this approach is non-decreasing and must have a limit. We remark that since $G(\boldsymbol{\lambda},\boldsymbol{\mu})$ is jointly concave with respect to $(\boldsymbol{\lambda},\boldsymbol{\mu})$, the alternating minimization approach converges to the global optima [13, Proposition 2.7.1], i.e.,

$$\lim_{k\to\infty} G(\boldsymbol{\lambda}^{(k)},\boldsymbol{\mu}^{(k)}) = \lim_{k\to\infty} G(\boldsymbol{\lambda}^{(k+1)},\boldsymbol{\mu}^{(k)}) = L.$$

Next, we derive the expression of $\boldsymbol{\lambda}^{(k)}$ and $\boldsymbol{\mu}^{(k)}$ obtained from the alternating minimization procedure. Initially we set $\lambda_x^{(0)} = 0,\ \forall x \in \mathcal{X}$. For $k \geq 0$, by the definition of $\boldsymbol{\mu}^{(k)}$ in the alternating minimization we have $\frac{\partial G}{\partial \mu_y^{(k)}} = 0,\ \forall y \in \mathcal{Y}$, which gives

$$e^{-\mu_y^{(k)}} = \sum_{x:x\sim y} P_X(x) e^{\lambda_x^{(k)}}, \qquad \forall y \in \mathcal{Y}. \tag{28}$$

Similarly, for $\boldsymbol{\lambda}^{(k+1)}$ we have

$$e^{-\lambda_x^{(k+1)}} = \sum_{y:y\sim x} P_Y(y) e^{\mu_y^{(k)}}, \qquad \forall x \in \mathcal{X}. \tag{29}$$

Based on (29) and (28), it is straightforward to verify that the first two iterations of this procedure yield

$$\begin{aligned} L &\geq G\left(\boldsymbol{\lambda}^{(0)}, \boldsymbol{\mu}^{(0)}\right) = -\mathbb{E}_Y \log \mathbb{E}_X \mathbb{1}(X \sim Y) \\ L &\geq G\left(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\mu}^{(0)}\right) = -\mathbb{E}_Y \log \mathbb{E}_X \mathbb{1}(X \sim Y) \\ &\qquad - \mathbb{E}_X \log \mathbb{E}_Y \frac{\mathbb{1}(X \sim Y)}{\mathbb{E}_X \mathbb{1}(X \sim Y)} \end{aligned}$$

in agreement with the bound derived in Theorem 1. Continuing with this procedure we can further improve our bound. To characterize the bound after k iterations, we introduce the functions $T_X^{(k)}(P_X(x), P_Y(y), \mathbb{1}(x \sim y))$, $T_Y^{(k)}(P_X(x), P_Y(y), \mathbb{1}(x \sim y))$ that are defined recursively as

$$T_X^{(0)}(P_X(x), P_Y(y), \mathbb{1}(x \sim y)) = 1, \tag{30}$$

and for $k \geq 0$,

$$\begin{aligned} &T_Y^{(k)}(P_X(x), P_Y(y), \mathbb{1}(x \sim y)) \\ &\quad = \mathbb{E}_X \left(\frac{\mathbb{1}(X \sim Y)}{T_X^{(k)}(P_X(x), P_Y(y), \mathbb{1}(x \sim y))} \right), \qquad (31) \\ &T_X^{(k+1)}(P_X(x), P_Y(y), \mathbb{1}(x \sim y)) \\ &\quad = \mathbb{E}_Y \left(\frac{\mathbb{1}(X \sim Y)}{T_Y^{(k)}(P_X(x), P_Y(y), \mathbb{1}(x \sim y))} \right). \qquad (32) \end{aligned}$$

It can be easily verified by induction that

$$\begin{aligned} G\left(\boldsymbol{\lambda}^{(k)}, \boldsymbol{\mu}^{(k)}\right) &= -\mathbb{E}_X \log T_X^{(k)}(P_X(x), P_Y(y), \mathbb{1}(x \sim y)) \\ &\quad - \mathbb{E}_Y \log T_Y^{(k)}(P_X(x), P_Y(y), \mathbb{1}(x \sim y)) \\ G\left(\boldsymbol{\lambda}^{(k+1)}, \boldsymbol{\mu}^{(k)}\right) &= -\mathbb{E}_X \log T_X^{(k+1)}(P_X(x), P_Y(y), \mathbb{1}(x \sim y)) \\ &\quad - \mathbb{E}_Y \log T_Y^{(k)}(P_X(x), P_Y(y), \mathbb{1}(x \sim y)). \end{aligned}$$

Thus, we have arrived at the following theorem.

Theorem 4: For any jointly distributed discrete r.vs (X, Y) and any $k \geq 0$,

$$\begin{aligned} I(X;Y) &\geq -\mathbb{E}_X \log T_X^{(k+1)}(P_X(x), P_Y(y), \mathbb{1}(x \sim y)) \\ &\quad - \mathbb{E}_Y \log T_Y^{(k)}(P_X(x), P_Y(y), \mathbb{1}(x \sim y)) \\ &\geq -\mathbb{E}_X \log T_X^{(k)}(P_X(x), P_Y(y), \mathbb{1}(x \sim y)) \\ &\quad - \mathbb{E}_Y \log T_Y^{(k)}(P_X(x), P_Y(y), \mathbb{1}(x \sim y)). \end{aligned}$$

V. Examples

In this section we evaluate the bounds derived in Theorems 1 and 2, and when possible also those from Theorem 4, for four examples. The following simple lower bound on $I(X;Y)$ will serve as our baseline for demonstrating the improvement attained by applying the bound from Theorem 1.

Proposition 2: For any jointly distributed discrete r.vs (X, Y),

$$I(X, Y) \geq -\mathbb{E}_Y \log \mathbb{E}_X \mathbb{1}(X \sim Y). \tag{33}$$

Similar to the bound from Theorem 1, the bound above is given in terms of the marginals and the joint support of (X, Y). However it is weaker than the former bound as it can be obtained from it directly by applying Jensen's inequality on the second term of (1), which gives $-\mathbb{E}_X \log \mathbb{E}_Y \frac{\mathbb{1}(X\sim Y)}{\mathbb{E}_X \mathbb{1}(X\sim Y)} \geq -\log \mathbb{E}_Y \frac{\mathbb{E}_X \mathbb{1}(X\sim Y)}{\mathbb{E}_X \mathbb{1}(X\sim Y)} = 0$. In Section I we also gave an operational proof of this bound.

A. Erasure Channel

The binary erasure channel has input $X \in \{0, 1\}$ and output $Y \in \{0, 1, \mathcal{E}\}$ such that $\Pr(Y = x | X = x) = 1 - \epsilon$ and $\Pr(Y = \mathcal{E} | X = x) = \epsilon$ for any x. For $X \sim \text{Bern}(p)$ we have $\Pr(Y = 0) = (1-\epsilon)(1-p)$, $\Pr(Y = 1) = (1-\epsilon)p$ and $\Pr(Y = \mathcal{E}) = \epsilon$ and the mutual information between the input and output is $I_p(X;Y) = (1-\epsilon)h(p)$. For this channel $x \sim y$ if and only if either $x = y$ or $y = \mathcal{E}$, and therefore

$$\begin{aligned} -\mathbb{E}_Y \log \mathbb{E}_X \mathbb{1}(X \sim Y) &= -(1-\epsilon)(1-p)\log(1-p) \\ &\quad - (1-\epsilon)p\log(p) - \epsilon\log(1) \\ &= (1-\epsilon)h(p). \end{aligned}$$

Thus, for this channel our lower bound from Theorem 1 as well as the weaker bound from Proposition 2 are tight.

In order to evaluate our upper bound from Theorem 2 we need to choose an action A consistent with $P_{Y|X}$. We take the natural action set, that consists of two actions, $a_1(x) = x$ and $a_2(x) = \mathcal{E}$ with $p(a_1) = 1 - \epsilon$ and $p(a_2) = \epsilon$. For this choice we have

$$\begin{aligned} \mathbb{E}_A \log \mathbb{E}_{XY} I(A \sim (X, Y)) &= p(a_1)\log\Pr(X = Y) \\ &\quad + p(a_2)\log\Pr(Y = \mathcal{E}) \\ &= (1-\epsilon)\log(1-\epsilon) + \epsilon\log(\epsilon) \\ &= -h(\epsilon). \end{aligned}$$

Since $H(Y) = h(\epsilon) + (1-\epsilon)h(p)$, the upper bound from Theorem 2 is tight and gives

$$I_p(X;Y) \leq (1-\epsilon)h(p).$$

B. Z Channel

The (symmetric) Z channel has a binary input X and a binary output Y such that $\Pr(Y = 0 | X = 0) = 1$ and $\Pr(Y = 0 | X = 1) = \Pr(Y = 1 | X = 1) = \frac{1}{2}$. For $X \sim \text{Bern}(p)$ we have $Y \sim \text{Bern}(\frac{p}{2})$, and the mutual information between the input and output is $I_p(X;Y) = h(\frac{p}{2}) - p$. For this channel $x \sim y$ if and only if $(x, y) \neq (0, 1)$ and therefore

$\mathbb{E}_X \mathbb{1}(X \sim 0) = 1$ and $\mathbb{E}_X \mathbb{1}(X \sim 1) = \Pr(X = 1) = p$. We have

$$\begin{aligned} -\mathbb{E}_Y \log \mathbb{E}_X \mathbb{1}(X \sim Y) &= -\Pr(Y=0) \log \mathbb{E}_X \mathbb{1}(X \sim 0) \\ &\quad - \Pr(Y=1) \log \mathbb{E}_X \mathbb{1}(X \sim 1) \\ &= -\frac{p}{2} \log(p), \end{aligned} \tag{34}$$

and

$$\begin{aligned} &-\mathbb{E}_X \log \mathbb{E}_Y \frac{\mathbb{1}(X \sim Y)}{\mathbb{E}_X \mathbb{1}(X \sim Y)} \\ &\quad = -(1-p) \log \left(\left(1 - \frac{p}{2}\right) \frac{\mathbb{1}(0 \sim 0)}{1} + \frac{p}{2} \frac{\mathbb{1}(0 \sim 1)}{p} \right) \\ &\qquad - p \log \left(\left(1 - \frac{p}{2}\right) \frac{\mathbb{1}(1 \sim 0)}{1} + \frac{p}{2} \frac{\mathbb{1}(1 \sim 1)}{p} \right) \\ &\quad = -(1-p) \log \left(1 - \frac{p}{2}\right) - p \log \left(\left(1 - \frac{p}{2}\right) + \frac{1}{2} \right) \\ &\quad = 1 - (1-p) \log(2-p) - p \log(3-p). \end{aligned} \tag{35}$$

Thus, Proposition 2 gives

$$I_p(X;Y) \geq -\frac{p}{2} \log(p), \tag{36}$$

and Theorem 1 gives

$$\begin{aligned} I_p(X;Y) \geq 1 - \frac{p}{2} \log(p) - (1-p) \log(2-p) \\ - p \log(3-p). \end{aligned} \tag{37}$$

For comparison, we also take a look at a further refinement given by Theorem 4. By the definitions of $T_X^{(k)}$ and $T_Y^{(k)}$, we know that

$$\begin{aligned} &T_Y^{(0)}(P_X(x), P_Y(y), \mathbb{1}(x \sim y)) \\ &\quad = \mathbb{E}_X \mathbb{1}(X \sim Y) \\ &\quad = \mathbb{1}(Y = 0) + p\mathbb{1}(Y = 1), \\ &T_X^{(1)}(P_X(x), P_Y(y), \mathbb{1}(x \sim y)) \\ &\quad = \mathbb{E}_Y \left(\frac{\mathbb{1}(X \sim Y)}{T_Y^{(0)}(P_X(x), P_Y(y), \mathbb{1}(x \sim y))} \right) \\ &\quad = \mathbb{E}_Y \left(\frac{\mathbb{1}(X \sim Y)}{\mathbb{1}(Y=0) + p\mathbb{1}(Y=1)} \right) \\ &\quad = \left(1 - \frac{p}{2} + \frac{p}{2} \cdot 0\right) \mathbb{1}(X = 0) \\ &\qquad + \left(1 - \frac{p}{2} + \frac{p}{2} \cdot \frac{1}{p}\right) \mathbb{1}(X = 1) \\ &\quad = \left(1 - \frac{p}{2}\right) \mathbb{1}(X=0) + \left(\frac{3}{2} - \frac{p}{2}\right) \mathbb{1}(X=1), \\ &T_Y^{(1)}(P_X(x), P_Y(y), \mathbb{1}(x \sim y)) \\ &\quad = \mathbb{E}_X \left(\frac{\mathbb{1}(X \sim Y)}{T_X^{(1)}(P_X(x), P_Y(y), \mathbb{1}(x \sim y))} \right) \\ &\quad = \mathbb{E}_X \left(\frac{\mathbb{1}(X \sim Y)}{\left(1 - \frac{p}{2}\right) \mathbb{1}(X=0) + \left(\frac{3}{2} - \frac{p}{2}\right) \mathbb{1}(X=1)} \right) \\ &\quad = \left((1-p) \cdot \frac{1}{1 - \frac{p}{2}} + p \cdot \frac{1}{\frac{3}{2} - \frac{p}{2}} \right) \mathbb{1}(Y=0) \\ &\qquad + \left((1-p) \cdot 0 + p \cdot \frac{1}{\frac{3}{2} - \frac{p}{2}} \right) \mathbb{1}(Y=1) \\ &\quad = \left(\frac{2-2p}{2-p} + \frac{2p}{3-p} \right) \mathbb{1}(Y=0) + \frac{2p}{3-p} \mathbb{1}(Y=1). \end{aligned}$$

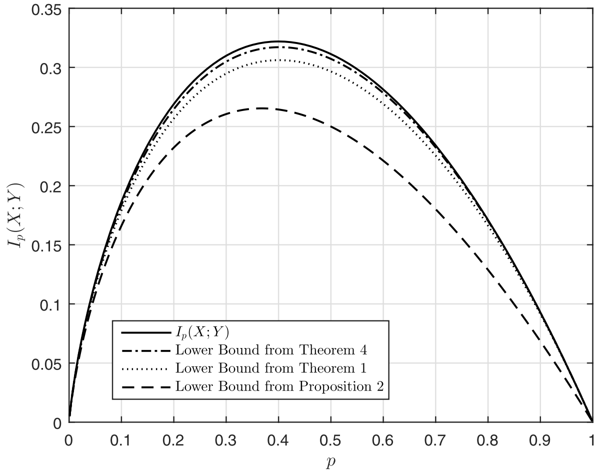

Fig. 1. $I_p(X;Y)$ for the Z channel together with the three lower bounds from (36), (37) and (38) as a function of p.

As a result, Theorem 4 gives

$$\begin{aligned} I_p(X;Y) &\geq -\mathbb{E}_X \log T_X^{(1)}(P_X(x), P_Y(y), \mathbb{1}(x \sim y)) \\ &\quad - \mathbb{E}_Y \log T_Y^{(1)}(P_X(x), P_Y(y), \mathbb{1}(x \sim y)) \\ &= -(1-p) \log \left(1 - \frac{p}{2}\right) - p \log \left(\frac{3-p}{2}\right) \\ &\quad - \left(1 - \frac{p}{2}\right) \log \left(\frac{2-2p}{2-p} + \frac{2p}{3-p} \right) \\ &\quad - \frac{p}{2} \log \left(\frac{2p}{3-p} \right) \\ &= \frac{p}{2} \log(2-p) + (1-p) \log(3-p) \\ &\quad - \frac{p}{2} \log(p) - \left(1 - \frac{p}{2}\right) \log(3-2p). \end{aligned} \tag{38}$$

The bounds from (36), (37) and (38) are plotted in Figure 1 as a function of p along with the exact value of $I_p(X;Y)$. It can be seen that the lower bound from Theorem 1 is significantly tighter than the one form Proposition 2, and it is quite close to $I_p(X;Y)$ for all values of p. The lower bound from Theorem 4 is even tighter.

In order to evaluate the upper bound from Theorem 2 we use the natural action $a_1(x) = x$ and $a_2(x) = 0$ with $p(a_1) = p(a_2) = \frac{1}{2}$. For this choice $a_1 \sim (x, y)$ if and only if $x = y$ and $a_2 \sim (x, y)$ if and only if $y = 0$, and therefore $\mathbb{E}_{XY} \mathbb{1}(a_1 \sim (X,Y)) = \Pr(X = Y) = 1 - \frac{p}{2}$ and $\mathbb{E}_{XY} \mathbb{1}(a_2 \sim (X,Y)) = \Pr(Y = 0) = 1 - \frac{p}{2}$. We have

$$\mathbb{E}_A \log \mathbb{E}_{XY} \mathbb{1}(A \sim (X,Y)) = \log \left(1 - \frac{p}{2}\right), \tag{39}$$

and

$$\begin{aligned} &\mathbb{E}_{X,Y} \log \mathbb{E}_A \frac{\mathbb{1}(A \sim (X,Y))}{\mathbb{E}_{X,Y} \mathbb{1}(A \sim (X,Y))} \\ &\quad = \mathbb{E}_{X,Y} \log \mathbb{E}_A \mathbb{1}(A \sim (X,Y)) - \log \left(1 - \frac{p}{2}\right) \\ &\quad = \Pr(X=0, Y=0) \log(1) + \Pr(X=1, Y=0) \log \left(\frac{1}{2}\right) \\ &\qquad + \Pr(X=1, Y=1) \log \left(\frac{1}{2}\right) - \log \left(1 - \frac{p}{2}\right) \\ &\quad = -p - \log \left(1 - \frac{p}{2}\right). \end{aligned} \tag{40}$$

Recalling that $H(Y) = h(\frac{p}{2})$ and applying theorem 2 we get

$$\begin{aligned} I_p(X;Y) &\le h\left(\frac{p}{2}\right) + \log\left(1-\frac{p}{2}\right) - p - \log\left(1-\frac{p}{2}\right) \\ &= h\left(\frac{p}{2}\right) - p, \end{aligned}$$

which is tight for any p.

C. Binary Deletion Channel

The binary i.i.d. deletion channel operates by independently deleting input bits with probability d. In this subsection, we apply Theorem 1 and Theorem 2 to obtain lower and upper bounds on the mutual information for an i.i.d. uniform input process. Both bounds outperform the best known bounds in some regimes of deletion probabilities. In general, tighter lower bounds can be obtained by applying Theorem 4 with higher values of k. However, as will be demonstrated below, even the task of computing the bound from Theorem 1 (corresponding to Theorem 4 with $k = 0$) is quite challenging.

1) Lower Bound for an i.i.d Uniform Input: We apply Theorem 1 to obtain a lower bound for $I(\mathbf{X};\mathbf{Y})$ under a uniform i.i.d. input distribution $\mathbf{X} \sim \text{Unif}(\{0,1\}^n)$, which outperforms the best known bounds for i.i.d inputs [14], [15]. Since the deletion channel is information stable, any rate smaller than the associated $\lim_{n\to\infty} I(\mathbf{X};\mathbf{Y})/n$ is achievable with uniform i.i.d. codebooks. Note that for a uniform i.i.d. input, the output $\mathbf{Y}$ is also uniform i.i.d. given its length Θn, where the latter is binomial with parameters $(n, 1-d)$.

For the i.i.d. deletion channel $\mathbb{1}(\mathbf{x}\sim\mathbf{y})$ indicates whether or not $\mathbf{y}$ is a subsequence of $\mathbf{x}$. For $0 \le t \le 1$, define the operation $\langle t\rangle \triangleq \max(t, 1/2)$. According to [3, Lemma 3.1], for any $\mathbf{y}$ of length θn we have

$$\sum_{\mathbf{x}\in\{0,1\}^n} \mathbb{1}(\mathbf{x}\sim\mathbf{y}) = \sum_{j=\theta n}^{n} \binom{n}{j} \doteq 2^{nh(\langle\theta\rangle)}, \tag{41}$$

where $h(\cdot)$ is the binary entropy function, and $\doteq$ denotes exponential equality in the usual sense. This implies that for any $\mathbf{y}$ of length θn we have $\mathbb{E}_{\mathbf{X}}\mathbb{1}(\mathbf{X}\sim\mathbf{y}) \doteq 2^{n(h(\langle\theta\rangle)-1)}$. The function $h(\langle\theta\rangle)$ is concave in θ and therefore

$$\begin{aligned} -\lim_{n\to\infty}\frac{1}{n}\mathbb{E}_Y \log \mathbb{E}_X \mathbb{1}(\mathbf{X}\sim\mathbf{Y}) &= -\mathbb{E}_\Theta\left(h(\langle\Theta\rangle) - 1\right) \\ &\ge 1 - h(\langle\mathbb{E}\Theta\rangle) \\ &= 1 - h(\langle 1-d\rangle). \end{aligned} \tag{42}$$

where Θ is the normalized (random) length of Y.

The right hand side of (42) is a well known lower bound for the deletion channel capacity, obtained with a uniform i.i.d. input [3]. We now evaluate the second term in (1) in order to improve upon this bound. To this end, we first parse each $x \in \{0,1\}^n$ into phrases that contain exactly two bit flips and end immediately after the second flip. For example, the string 0001111011001110001 is parsed into the three phrases 00011110, 11001, 110001. We identify each phrase with three parameters: $b \in \{0,1\}$ is the first bit in the phrase, $k_1 \ge 2$ is the index of the first flip in the phrase, and $k_2 \ge 1$ is such that $k_1 + k_2$ is the total number of bits in the phrase. In our example, the three phrases correspond to $\{b = 0, k_1 = 4, k_2 = 4\}$, $\{b = 1, k_1 = 3, k_2 = 2\}$ and $\{b = 1, k_1 = 3, k_2 = 3\}$, respectively. For any pair of integers $2 \le k_1 < n$, $1 \le k_2 < n$ let $\Psi^{k_1,k_2}(\mathbf{x})$ be the number of $\{k_1,k_2\}$-phrases in the parsing of $\mathbf{x}$. For $\epsilon > 0$ we define the typical set

$$\mathcal{S}_\epsilon \triangleq \left\{\mathbf{x}\in\{0,1\}^n \;:\; \left|\frac{1}{n}\Psi^{k_1,k_2}(\mathbf{x}) - \frac{1}{5}\cdot 2^{-(k_1+k_2-1)}\right| < \epsilon \quad \forall\, 2 \le k_1 < n,\; 1 \le k_2 < n\right\}.$$

It holds that for any $\epsilon > 0$ and n large enough $\Pr(\mathbf{X}\in\mathcal{S}_\epsilon)$ is indeed arbitrary close to 1. To see this, define the three i.i.d. mutually independent processes

$$\begin{aligned} B_i &\sim \text{Bern}(\tfrac{1}{2}),\ \text{i.i.d.} \\ K_{1i} &\sim 1 + \text{Geometric}(\tfrac{1}{2}),\ \text{i.i.d.} \\ K_{2i} &\sim \text{Geometric}(\tfrac{1}{2}),\ \text{i.i.d.} \end{aligned}$$

and note that an i.i.d. $\text{Bern}(\frac{1}{2})$ random process is equivalent to the process obtained by stacking the random phrases $\{B_i, K_{1i}, K_{2i}\}$ one after the other. Moreover, the probability of such a random phrase being of type $\{k_1,k_2\}$ is $2^{-(k_1+k_2-1)}$ and the expected length is $\mathbb{E}(K_{1i}+K_{2i}) = 5$. In our setting, $\mathbf{X}$ is an n-dimensional i.i.d. $\text{Bern}(\frac{1}{2})$ random vector. Thus, $\mathbf{X}$ can be generated by stacking exactly $n/5$ random phrases $\{B_i, K_{1i}, K_{2i}\}$ one after the other and either removing the last bits if the length of the obtained vector is greater than n, or appending i.i.d. $\text{Bern}(\frac{1}{2})$ bits to the vector if its length is smaller than n. Since the expected length of a phrase is 5 bits, for any $\delta > 0$ the number of removed/appended bits is w.h.p. smaller than δn. Therefore, the contribution of these bits to the distribution of the phrase lengths in the parsing of $\mathbf{X}$ is negligible, and we get that $\Pr(\mathbf{X}\in\mathcal{S}_\epsilon) \to 1$ with n, by the law of large numbers.

For n large enough we can write

$$\begin{aligned} &-\mathbb{E}_{\mathbf{X}} \log \mathbb{E}_{\mathbf{Y}} \frac{\mathbb{1}(\mathbf{X}\sim\mathbf{Y})}{\mathbb{E}_{\mathbf{X}}\mathbb{1}(\mathbf{X}\sim\mathbf{Y})} \\ &\quad= -\Pr(\mathbf{X}\in\mathcal{S}_\epsilon)\mathbb{E}_{\mathbf{X}|\mathcal{S}_\epsilon} \log \mathbb{E}_{\mathbf{Y}} \frac{\mathbb{1}(\mathbf{X}\sim\mathbf{Y})}{\mathbb{E}_{\mathbf{X}}\mathbb{1}(\mathbf{X}\sim\mathbf{Y})} \\ &\qquad - \Pr(\mathbf{X}\notin\mathcal{S}_\epsilon)\mathbb{E}_{\mathbf{X}|\overline{\mathcal{S}}_\epsilon} \log \mathbb{E}_{\mathbf{Y}} \frac{\mathbb{1}(\mathbf{X}\sim\mathbf{Y})}{\mathbb{E}_{\mathbf{X}}\mathbb{1}(\mathbf{X}\sim\mathbf{Y})} \\ &\quad\ge -\Pr(\mathbf{X}\in\mathcal{S}_\epsilon) \log \mathbb{E}_{\mathbf{Y}} \frac{\mathbb{E}_{\mathbf{X}|\mathcal{S}_\epsilon}\mathbb{1}(\mathbf{X}\sim\mathbf{Y})}{\mathbb{E}_{\mathbf{X}}\mathbb{1}(\mathbf{X}\sim\mathbf{Y})} \\ &\qquad - \Pr(\mathbf{X}\notin\mathcal{S}_\epsilon) \log \mathbb{E}_{\mathbf{Y}} \frac{\mathbb{E}_{\mathbf{X}|\overline{\mathcal{S}}_\epsilon}\mathbb{1}(\mathbf{X}\sim\mathbf{Y})}{\mathbb{E}_{\mathbf{X}}\mathbb{1}(\mathbf{X}\sim\mathbf{Y})} \\ &\quad\ge -(1-\epsilon) \log \mathbb{E}_{\mathbf{Y}} \frac{\mathbb{E}_{\mathbf{X}|\mathcal{S}_\epsilon}\mathbb{1}(\mathbf{X}\sim\mathbf{Y})}{\mathbb{E}_{\mathbf{X}}\mathbb{1}(\mathbf{X}\sim\mathbf{Y})} - \epsilon n, \end{aligned} \tag{43}$$

where the first inequality follows from Jensen's inequality and in the second we have used the fact that $\mathbb{E}_X\mathbb{1}(\mathbf{X}\sim\mathbf{y}) \ge 2^{-n}$ for any $\mathbf{y}$, and therefore $\mathbb{1}(\mathbf{X}\sim\mathbf{Y})/\mathbb{E}_X\mathbb{1}(\mathbf{X}\sim\mathbf{y}) \le 2^n$ for any $\mathbf{y}$, along with $\Pr(\mathbf{X}\in\mathcal{S}_\epsilon) > 1-\epsilon$. Recalling that Θ is the normalized (random) length of $\mathbf{Y}$, we take the expectation $\mathbb{E}_{\mathbf{Y}}$ as $\mathbb{E}_\Theta\mathbb{E}_{\mathbf{Y}|\Theta}$ and use (41) to obtain

$$\begin{aligned} &\mathbb{E}_{\mathbf{Y}} \frac{\mathbb{E}_{\mathbf{X}|\mathcal{S}_\epsilon}\mathbb{1}(\mathbf{X}\sim\mathbf{Y})}{\mathbb{E}_{\mathbf{X}}\mathbb{1}(\mathbf{X}\sim\mathbf{Y})} \\ &\quad\doteq \mathbb{E}_\Theta 2^{n(1-h(\langle\Theta\rangle))}\mathbb{E}_{\mathbf{Y}|\Theta}\mathbb{E}_{\mathbf{X}|\mathcal{S}_\epsilon}\mathbb{1}(\mathbf{X}\sim\mathbf{Y}) \\ &\quad= \mathbb{E}_\Theta 2^{n(1-h(\langle\Theta\rangle))} \Pr\left(\mathbf{X}\sim\mathbf{Y}|\Theta, \mathbf{X}\in\mathcal{S}_\epsilon\right). \end{aligned} \tag{44}$$

Now, consider a greedy algorithm for determining whether $\mathbf{y}$ is a subsequence of $\mathbf{x}$, defined as follows [2, Sec. 3.1]: Scanning from left to right, take the first bit in $\mathbf{y}$ and match it with its first appearance in $\mathbf{x}$. Then take the second bit in $\mathbf{y}$ and match it with its subsequent first appearance in $\mathbf{x}$. Continue until either $\mathbf{x}$ or $\mathbf{y}$ are exhausted, where the latter case is termed success. It is easy to see that the greedy algorithm succeeds if and only if $\mathbf{x} \sim \mathbf{y}$. For statistically independent random vectors $\mathbf{X}$ and $\mathbf{Y}$, we enumerate the phrases in the parsing of $\mathbf{X}$ by $i = 1, \dots, M(\mathbf{X})$ where $M(\mathbf{X})$ is the (random) number of phrases in $\mathbf{X}$. The vector $\mathbf{Y}$ consists of Θn i.i.d. uniform bits. To simplify computations, we construct a vector $\mathbf{Y}'$ of length n by taking $\mathbf{Y}$ and possibly padding it with i.i.d. bits. We define the random variables Z_i as the number of bits in $\mathbf{Y}'$ that are matched to bits in the ith phrase of $\mathbf{X}$ by the greedy algorithm. Under this construction, the events $\left\{\sum_i Z_i \geq \Theta n\right\}$ coincides with the event $\{\mathbf{X} \sim \mathbf{Y}\}$, since the additional random suffix does not affect the event where the first Θn bits in $\mathbf{Y}'$ are matched. Under this assumption the Z_i's are clearly mutually independent, given that the phrase types $\{k_{1i}, k_{2i}\}_{i=1}^{M(\mathbf{X})}$ of $\mathbf{X}$ are known (but assuming that their first bit identifiers $\{b_i\}_{i=1}^{M(\mathbf{X})}$ remain random). Of course, the distribution of Z_i depends on the parameters k_{1i}, k_{2i} that correspond to the ith phrase in $\mathbf{X}$. In the appendix, we show that given K_{1i} and K_{2i}, the (base two) moment generating function of Z_i is

$$\lambda_{Z_i}^{k_1,k_2}(t) \triangleq \mathbb{E}\left(2^{tZ_i}|K_{1i}=k_1, K_{2i}=k_2\right) = 2^{k_1(t-1)} + 2^{t-1}\frac{1-2^{k_1(t-1)}}{1-2^{t-1}} \times \left(2^{t-1}\frac{1-2^{k_2(t-1)}}{1-2^{t-1}} + 2^{k_2(t-1)-t}\right).$$

Noting that by definition, for $\mathbf{X} \in \mathcal{S}_\epsilon$ the number of phrases $M(\mathbf{X})$ and their composition $\Psi^{k_1,k_2}(\mathbf{X})$ is essentially deterministic, we can use Chernoff's bound [16] to obtain

$$\Pr(\mathbf{X} \sim \mathbf{Y}|\Theta = \theta, \mathbf{X} \in \mathcal{S}_\epsilon) = \Pr\left(\sum_{i=1}^{M(\mathbf{X})} Z_i \geq \Theta n|\Theta = \theta, \mathbf{X} \in \mathcal{S}_\epsilon\right) \dot{\leq} 2^{-n\Lambda^*(\theta)},$$

where

$$\Lambda^*(\theta) = \max_{t>0}\left(\theta t - \frac{1}{5}\sum_{k_1=2}^{\infty}\sum_{k_2=1}^{\infty} 2^{-(k_1+k_2-1)} \log \lambda_{Z_{k_1,k_2}}(t)\right).$$

Substituting into (43) and (44), and applying standard large deviations arguments, we obtain

$$-\lim_{n\to\infty}\frac{1}{n}\mathbb{E}_{\mathbf{X}} \log \mathbb{E}_{\mathbf{Y}} \frac{\mathbb{1}(\mathbf{X} \sim \mathbf{Y})}{\mathbb{E}_{\mathbf{X}}\mathbb{1}(\mathbf{X} \sim \mathbf{Y})} \geq g(d)$$

where

$$g(d) \triangleq \min_{0\leq\theta\leq1} D_2(\theta\|1-d) - (1 - h(\langle\theta\rangle)) + \Lambda^*(\theta)$$

where $D_2(p\|q)$ is the binary relative entropy function. It follows that for a uniform i.i.d. input distribution,

$$\lim_{n\to\infty}\frac{1}{n}I(\mathbf{X};\mathbf{Y}) \geq 1 - h(\min(d, 1/2)) + g(d). \tag{45}$$

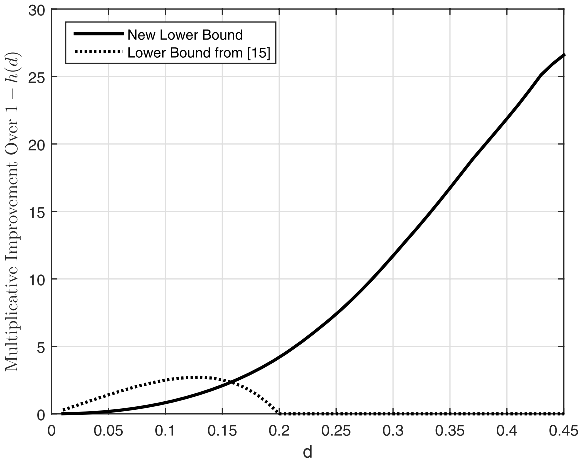

Fig. 2. The multiplicative improvement factor w.r.t. $1 - h(d)$ attained by our lower bound on the mutual information for an i.i.d. uniform input. For comparison, we also plot the improvement the lower bound from [15] attains w.r.t. $1 - h(d)$.

Numerical evaluation of the term $g(d)$ reveals that it is greater than zero for all $d < 1/2$. Thus, (45) improves over Gallager's well know bound $1 - h(d)$ [14]. Recently, Rahmati and Duman [15] used a different technique to lower bound the mutual information for uniform i.i.d. inputs. For small values of d their bound is better than (45), but for larger values of d the right hand side of (45) turns out to be greater than their bound. For example, for $d = 0.2$ our bound improves on $1-h(0.2)$ by ≈ 0.0117 bits (roughly 5%), whereas the improvement of [15] is negligible. See Figure 2.

2) Upper Bound for i.i.d Inputs: By Theorem 2 we have in particular that

$$I(\mathbf{X};\mathbf{Y}) \leq H(\mathbf{Y}) + \mathbb{E}_A \log \mathbb{E}_{\mathbf{X},\mathbf{Y}}\mathbb{1}(A \sim (\mathbf{X},\mathbf{Y})) \tag{46}$$

Let $\mathbf{X}$ be an i.i.d. Bern(q) input vector of length n for some $q \leq \frac{1}{2}$. It can be shown that the length of $\mathbf{Y}$ is $\Theta \sim$ Binomial($n, 1-d$), and given its length, $\mathbf{Y}$ is i.i.d. Bern(q). Thus,

$$\frac{1}{n}H(\mathbf{Y}) = \frac{1}{n}\left(H(\mathbf{Y}|\Theta) + H(\Theta)\right) = (1-d)h(q) + O\left(\frac{\log n}{n}\right) \tag{47}$$

The challenge is thus to evaluate the second term in (46), which is given by

$$\begin{aligned}\mathbb{E}_A \log \mathbb{E}_{\mathbf{X},\mathbf{Y}}\mathbb{1}(A \sim (\mathbf{X},\mathbf{Y})) &= \mathbb{E}_A \log \mathbb{E}_{A'}\mathbb{E}_{\mathbf{X}}\mathbb{1}(A \sim (\mathbf{X},A')) \\ &= \mathbb{E}_A \log \mathbb{E}_{A'}\mathbb{E}_{\mathbf{X}}\mathbb{1}(A(\mathbf{X}) = A'(\mathbf{X})) \\ &= \mathbb{E}_A \log \mathbb{E}_{A'} \Pr(A(\mathbf{X}) = A'(\mathbf{X}))\end{aligned} \tag{48}$$

where $A' \sim P_A$ such that $(\mathbf{X}, A'(\mathbf{X})) \sim P_{\mathbf{XY}}$. Note that here $\mathbf{X}, A, A'$ are mutually independent.

Let us specifically choose A as in Example 3, namely we identify A with a Bern($1-d$) i.i.d. vector of length n, and $A(\mathbf{X})$ corresponds to sampling $\mathbf{X}$ in the location chosen by that vector. Asymptotically, we can assume without loss of generality that both A and A' are drawn uniformly over vectors of weight $n(1-d)$. This follows since for any given weight of A, the inner expectation w.r.t. A' only increases by replacing the i.i.d. distribution with a uniform distribution over all vectors with the same weight. Furthermore, the outer expectation w.r.t. A is asymptotically dominated by the uniform distribution over vectors of weight $n(1-d)$.

Let us define S to be the action that chooses only the coordinates selected by A' but not by A. Let $\overline{S}$ be the complementary action (that chooses only the remaining coordinates). Given any A' and A, for any assignment of the values of $\mathbf{X}$ in the coordinates chosen by $\overline{S}$, there is either a unique assignment $\phi(\overline{S}(\mathbf{X}))$ of the values of $\mathbf{X}$ in the coordinates chosen by S that satisfies $A'(\mathbf{X}) = A(\mathbf{X})$, or there is none. In the latter case, we set $\phi(\overline{S}(\mathbf{X}))$ to an arbitrary value. Thus we can write

$$\begin{aligned}
\Pr(A(\mathbf{X}) &= A'(\mathbf{X})) \\
&= \Pr\left(\mathbf{X} \in \left\{\mathbf{x} \in \{0,1\}^n : \mathbb{1}(A'(\mathbf{x}) = A(\mathbf{x})\right\}\right) \\
&\leq \Pr\left(\mathbf{X} \in \left\{\mathbf{x} \in \{0,1\}^n : S(\mathbf{x}) = \phi(\overline{S}(\mathbf{x}))\right\}\right) \\
&= \Pr\left(S(\mathbf{X}) = \phi(\overline{S}(\mathbf{X}))\right) \\
&= \mathbb{E}\Pr\left(S(\mathbf{X}) = \phi(\overline{S}(\mathbf{X})) \mid \overline{S}(\mathbf{X})\right) \\
&\leq \mathbb{E} \max_{\mathbf{u}\in\{0,1\}^{|S|}} \Pr\left(S(\mathbf{X}) = \mathbf{u} \mid \overline{S}(\mathbf{X})\right) \\
&= \mathbb{E} \max_{\mathbf{u}\in\{0,1\}^{|S|}} \Pr\left(S(\mathbf{X}) = \mathbf{u}\right) \\
&= \max_{\mathbf{u}\in\{0,1\}^{|S|}} \Pr\left(S(\mathbf{X}) = \mathbf{u}\right) \\
&= (1-q)^{|S|}
\end{aligned}$$

Returning to (48) and using the above, we have

$$\mathbb{E}_A \log \mathbb{E}_{A'} \Pr(A(\mathbf{X}) = A'(\mathbf{X})) \leq \mathbb{E}_A \log \mathbb{E}_{A'} (1-q)^{|S|}$$

where the only randomness is in $|S|$, which is a deterministic function of A and A'. In particular, $|S|$ is the number of coordinates chosen by A' and not by A. Since A and A' were assumed to be uniformly distributed over constant weight vectors of weight $(1-d)n$, then simple counting arguments show that for every action a

$$\begin{aligned}
\Pr(|S| = \rho(1-d)n | A = a) &= \frac{\binom{(1-d)n}{(1-\rho)(1-d)n} \cdot \binom{dn}{\rho(1-d)n}}{\binom{n}{(1-d)n}} \\
&\doteq 2^{n\left((1-d)h(\rho) + d\cdot h\left(\rho\frac{1-d}{d}\right) - h(d)\right)}
\end{aligned}$$

Thus, maximizing over feasible values of ρ

$$\begin{aligned}
\lim_{n\to\infty} &\frac{1}{n}\mathbb{E}_A \log \mathbb{E}_{A'} \Pr(A(\mathbf{X}) = A'(\mathbf{X})) \\
&\leq \max_{0\leq\rho\leq\frac{d}{1-d}} (1-d)h(\rho) + d\cdot h\left(\rho\frac{1-d}{d}\right) \\
&\quad - h(d) + (1-d)\rho\log(1-q)
\end{aligned}$$

Plugging the above in (46) and using (47), we obtain the bound

$$\lim_{n\to\infty} \frac{1}{n} I(\mathbf{X};\mathbf{Y}) \leq (1-d)h(q) - h(d) + \max_{0\leq\rho\leq\frac{d}{1-d}} \Gamma(\rho)$$

where

$$\Gamma(\rho) \triangleq (1-d)\left(h(\rho) + \rho\log(1-q)\right) + d\cdot h\left(\rho\frac{1-d}{d}\right).$$

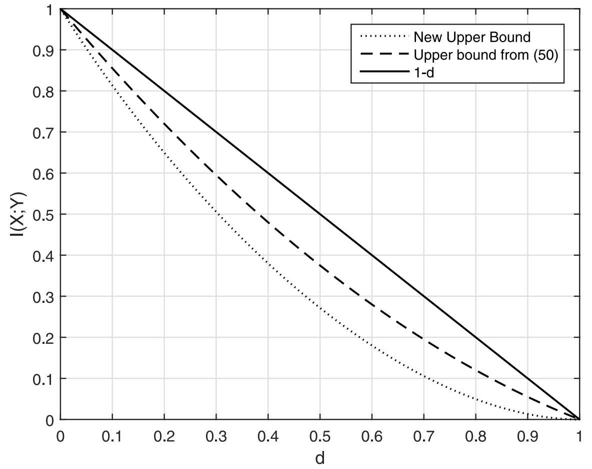

Fig. 3. Our new upper bound (49) plotted for $q = 1/2$ along with the upper bound (50) and the trivial upper bound $1-d$.

We note that the maximization over ρ can be solved directly by differentiation, and the maximizing value is

$$\rho^* = \frac{1-q}{2q(1-d)}\left(\sqrt{1 + 4d(1-d)\frac{q}{1-q}} - 1\right),$$

and we therefore have

$$\lim_{n\to\infty} \frac{1}{n} I(\mathbf{X};\mathbf{Y}) \leq (1-d)h(q) - h(d) + \Gamma(\rho^*). \tag{49}$$

In the limit of $d \to 1$ it is easy to see that $\rho^* \to d$, and direct substitution into (49) reveals that for $q = 1/2$ the upper bound is smaller than $(1-d)^2$ for large d. In [17] it was shown that for an i.i.d. Bern(q) input process

$$\lim_{n\to\infty} \frac{1}{n} I(\mathbf{X};\mathbf{Y}) \leq (1-d)\left(h(q) - 2dq(1-q)\right). \tag{50}$$

Our new upper bound is plotted in Figure 3 for $q = 1/2$ along with the upper bound (50) and the trivial upper bound $1-d$. It is seen that for this choice of q our new bound is better than (50) for all deletion probabilities.

We remark that although here we have only applied the bounds from Theorems 1 and 2 for handling deletion channels, we expect a similar approach to yield improved results also for insertion channels.

D. Most Informative Boolean Function Conjecture

Let $\mathbf{X}$ be an n-dimensional binary vector uniformly distributed over $\{0,1\}^n$, and $\mathbf{Y}$ be the output of passing each component of $\mathbf{X}$ through a binary symmetric channel with crossover probability $\alpha \leq 1/2$. Let $f : \{0,1\}^n \to \{0,1\}$ be a boolean function. Following a recent conjecture by Courtade and Kumar [6], there has been much interest in developing useful upper bounds on $I(f(\mathbf{X});\mathbf{Y})$, where the ultimate goal is to prove that this quantity is maximized by the dictatorship function $f(\mathbf{X}) = X_i$ for some $i \in [n]$. In this subsection, we apply Theorem 2 to derive the following novel upper bound.

Theorem 5: Let $\mathbf{X}, \mathbf{Z}, \mathbf{W} \in \{0,1\}^n$ be three statistically independent random vectors, with the entries of $\mathbf{X}$ i.i.d.

Bern($\frac{1}{2}$), and the entries of $\mathbf{Z}$ and $\mathbf{W}$ i.i.d. Bern(α). Let $\mathbf{Y} = \mathbf{X} \oplus \mathbf{Z}$. For any boolean function $f : \{0, 1\}^n \to \{0, 1\}$,

$$\begin{aligned} I(\mathbf{Y}; f(\mathbf{X})) &\leq H(f(\mathbf{X})) \\ &\quad + \mathbb{E}_{\mathbf{W}} \log \Pr(f(\mathbf{X} \oplus \mathbf{W} \oplus \mathbf{Z}) = f(\mathbf{X})) \end{aligned} \tag{51}$$

Proof: Identify the action that maps $\mathbf{Y}$ to $f(\mathbf{X})$ with drawing an i.i.d. vector $\mathbf{W}$ with Bern(α) entries and setting $A(\mathbf{Y}) = f(\mathbf{Y} \oplus \mathbf{W})$. The bound (3) reads (discarding the last term which is non-positive)

$$\begin{aligned} &I(\mathbf{Y}; f(\mathbf{X})) \\ &\leq H(f(\mathbf{X})) + \mathbb{E}_A \log \mathbb{E}_{\mathbf{Y}, f(\mathbf{X})} \mathbb{1}(A(\mathbf{Y}) = f(\mathbf{X})) \\ &= H(f(\mathbf{X})) + \mathbb{E}_{\mathbf{W}} \log \mathbb{E}_{\mathbf{Y}, f(\mathbf{X})} \mathbb{1}(f(\mathbf{Y} \oplus \mathbf{W}) = f(\mathbf{X})) \\ &= H(f(\mathbf{X})) + \mathbb{E}_{\mathbf{W}} \log \Pr(f(\mathbf{X} \oplus \mathbf{W} \oplus \mathbf{Z}) = f(\mathbf{X})), \end{aligned} \tag{52}$$

as desired. ■

For a fixed $\mathbf{w} \in \{0, 1\}^n$, let us now express $\Pr(f(\mathbf{X} \oplus \mathbf{w} \oplus \mathbf{Z}) = f(\mathbf{X}))$. To this end, we use the standard isomorphism $0 \to 1$, $1 \to -1$, $\oplus \to \cdot$. Under this isomorphism we need to calculate $\Pr(f(\mathbf{X} \cdot \mathbf{w} \cdot \mathbf{Z}) = f(\mathbf{X}))$, where the products between vectors are taken componentwise. Recall [18] that $f : \{-1, 1\}^n \to \{-1, 1\}$ admits the Fourier-Walsh expansion

$$f(\mathbf{x}) = \sum_{S \subseteq [n]} \hat{f}(S) \prod_{i \in S} x_i, \tag{53}$$

where

$$\hat{f}(S) \triangleq \mathbb{E}\left(f(\mathbf{X}) \prod_{i \in S} X_i \right), \tag{54}$$

and the expectation is taken w.r.t. to i.i.d. uniform distribution on $\{-1, 1\}$. Let $f_{\mathbf{w}}(\mathbf{X}) = f(\mathbf{X} \cdot \mathbf{w})$, and note that it immediately follows from (53) that $\hat{f}_{\mathbf{w}}(S) = \hat{f}(S) \prod_{i \in S} w_i$. We have

$$\begin{aligned} &\Pr(f(\mathbf{X} \cdot \mathbf{w} \cdot \mathbf{Z}) = f(\mathbf{X})) \\ &= \Pr(f_{\mathbf{w}}(\mathbf{X} \cdot \mathbf{Z}) = f(\mathbf{X})) \\ &= \frac{1}{2} \left(1 + \mathbb{E}\left(f(\mathbf{X}) f_{\mathbf{w}}(\mathbf{X} \cdot \mathbf{Z})\right)\right) \\ &= \frac{1}{2} \left(1 + \mathbb{E}\left(\sum_{S \subseteq [n]} \hat{f}(S) \prod_{i \in S} X_i \sum_{T \subseteq [n]} \hat{f}(T) \prod_{j \in T} X_j Z_j w_j \right) \right) \\ &= \frac{1}{2} \left(1 + \sum_{S \subseteq [n]} \hat{f}^2(S) (1 - 2\alpha)^{|S|} \prod_{i \in S} w_i \right), \end{aligned} \tag{55}$$

where in (55) we have used the facts that $\mathbb{E}(X_i X_j) = \mathbb{1}(i = j)$ and $\mathbb{E}(Z_i) = (1-2\alpha)$ for any $i, j \in [n]$. Now, substituting (55) into (52) gives the following corollary.

Corollary 1: For any boolean $f : \{-1, 1\}^n \to \{-1, 1\}$,

$$\begin{aligned} I(\mathbf{Y}; f(\mathbf{X})) &\leq H(f(\mathbf{X})) - 1 \\ &\quad + \mathbb{E}_{\mathbf{W}} \log \left(1 + \sum_{S \subseteq [n]} \hat{f}^2(S) (1 - 2\alpha)^{|S|} \prod_{i \in S} W_i \right). \end{aligned} \tag{56}$$

where W_i are i.i.d. with $\Pr(W_i = -1) = 1 - \Pr(W_i = 1) = \alpha$.

We note that the upper bound from Theorem 5 and Corollary 1 are tight for the function $f(\mathbf{X}) = X_i$. Thus, showing that the dictatorship function maximizes (52) or (56), will settle the most informative boolean function conjecture [6]. Unfortunately, our attempts to prove the former were not successful.

Appendix

Given that $K_{1i} = k_1$ and $K_{2i} = k_2$, we know that the ith phrase in the parsing of $\mathbf{X}$ is of the form

$$\underbrace{B \cdots B \overline{B}}_{k_1} \underbrace{\overline{B} \cdots \overline{B} B}_{k_2}, \tag{57}$$

where $B \sim$ Bern($\frac{1}{2}$) and $\overline{B} \triangleq 1 - B$. The r.v. Z_i counts the number of bits in $\mathbf{Y}'$ that were matched by the greedy algorithm to bits in the ith phrase of $\mathbf{X}$. Thus, conditioned on the event $K_{1i} = k_1, K_{2i} = k_2$, the r.v. Z_i counts the number of bits from an i.i.d. uniform sequence (corresponding to the relevant bits in $\mathbf{Y}'$) that are matched by the greedy algorithm to bits in the phrase (57).

Let W be the event that the first k_1 bits of the i.i.d. sequence are equal to B. Clearly, $\Pr(W) = 2^{-k_1}$ and if W occurs then $Z_i = k_1$. Let T_1 be the location of the first occurrence of $\overline{B}$ in the i.i.d. sequence, and let T_2' be the location of the first occurrence of B after T_1. Further, let $T_2 = T_2' - T_1$. For example, if the sequence of i.i.d. bits is

$$B \;\; B \overline{B} \;\; \overline{B} \;\; \overline{B} \;\; \overline{B} \;\; \overline{B} \;\; B \ldots,$$

then $T_1 = 3$ and $T_2 = 5$, and if the sequence of i.i.d. bits is

$$\overline{B} \;\; \overline{B} \;\; B \ldots,$$

then $T_1 = 1$ and $T_2 = 2$. We further define the r.v.

$$\tilde{T}_2 = \begin{cases} T_2 & T_2 \leq k_2 \\ k_2 - 1 & T_2 > k_2. \end{cases}$$

Note that given $\overline{W}$ (the event that W did not occur), we have $Z_i = T_1 + \tilde{T}_2$. We have

$$\begin{aligned} &\mathbb{E}\left(2^{tZ_i} \mid K_{1i} = k_1, K_{2i} = k_2\right) \\ &\quad = \Pr(W) \mathbb{E}\left(2^{tZ_i} \mid K_{1i} = k_1, K_{2i} = k_2, W\right) \\ &\qquad + \Pr(\overline{W}) \mathbb{E}\left(2^{tZ_i} \mid K_{1i} = k_1, K_{2i} = k_2, \overline{W}\right) \\ &\quad = 2^{-k_1} 2^{tk_1} + \left(1 - 2^{-k_1}\right) \mathbb{E}\left(2^{t(T_1 + \tilde{T}_2)} \mid \overline{W}\right) \\ &\quad = 2^{-k_1} 2^{tk_1} + \left(1 - 2^{-k_1}\right) \mathbb{E}\left(2^{tT_1} \mid \overline{W}\right) \mathbb{E}\left(2^{t\tilde{T}_2}\right). \end{aligned} \tag{58}$$

The r.v.s T_1 and T_2 are statistically independent Geometric($\frac{1}{2}$), and therefore

$$\Pr(T_1 = m | \overline{W}) = \begin{cases} \frac{2^{-m}}{1 - 2^{-k_1}} & 1 \leq m \leq k_1 - 1 \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\begin{aligned} &\Pr(\tilde{T}_2 = m) \\ &\quad = \begin{cases} 2^{-m} & 1 \leq m \leq k_2, m \neq k_2 - 1 \\ 2^{-k_2} + 2^{-m} \mathbb{1}(k_2 > 1) & m = k_2 - 1 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

This gives

$$\mathbb{E}\left(2^{tT_1}\right) = \frac{1}{1-2^{-k_1}}\sum_{m=1}^{k_1-1} 2^{-m}2^{tm}$$
$$= \frac{1}{1-2^{-k_1}}\frac{2^{t-1}}{1-2^{t-1}}\left(1-2^{k_1(t-1)}\right) \tag{59}$$

and for $k_2 > 1$

$$\mathbb{E}\left(2^{t\tilde{T}_2}\right) = \sum_{m=1}^{k_2} 2^{-m}2^{tm} + 2^{-k_2}2^{t(k_2-1)}$$
$$= \frac{2^{t-1}}{1-2^{t-1}}\left(1-2^{k_2(t-1)}\right) + 2^{k_2(t-1)-t}. \tag{60}$$

Note that for $k_2 = 1$ we have $\mathbb{E}\left(2^{t\tilde{T}_2}\right) = \frac{1}{2} + \frac{1}{2}2^{-t}$, and (60) continues to hold. Substituting (59) and (60) into (58) yields the desired expression.

References

[1] R. L. Dobrushin, "General formulation of Shannon's main theorem in information theory," *Amer. Math. Soc. Trans.*, vol. 33, no. 2, pp. 323–438, 1963.

[2] M. Mitzenmacher, "A survey of results for deletion channels and related synchronization channels," *Probab. Surv.*, vol. 6, no. 1, pp. 205–237, 2009.

[3] S. N. Diggavi and M. Grossglauser, "On transmission over deletion channels," in *Proc. Annu. Allerton Conf. Commun. Control Comput.*, 2001, vol. 39. no. 1, pp. 573–582.

[4] E. Drinea and M. Mitzenmacher, "On lower bounds for the capacity of deletion channels," *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4648–4657, Oct. 2006.

[5] A. Gamal and Y. Kim, *Network Information Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[6] T. A. Courtade and G. R. Kumar, "Which Boolean functions maximize mutual information on noisy inputs?" *IEEE Trans. Inf. Theory*, vol. 60, no. 8, pp. 4515–4525, Aug. 2014.

[7] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*. New York, NY, USA: Wiley, 1997.

[8] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 1991.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *J. Roy. Statist. Soc. Ser. B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977. s

[10] R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. 18, no. 4, pp. 460–473, Jul. 1972.

[11] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. 18, no. 1, pp. 14–20, Jan. 1972.

[12] I. Csiszár and G. Tusnády, "Information geometry and alternating minimization procedures," *Statist. Decisions*, pp. 205–237, 1984.

[13] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA, USA: Athena scientific, 1999.

[14] R. G. Gallager, "Sequential decoding for binary channels with noise and synchronization errors," Massachusetts INST of Tech Lexington Lincoln Lab, MA, USA, Tech. Rep. AD0266879, 1961.

[15] M. Rahmati and T. M. Duman, "Bounds on the capacity of random insertion and deletion-additive noise channels," *IEEE Trans. Inf. Theory*, vol. 59, no. 9, pp. 5534–5546, Sep. 2013.

[16] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Berlin, Germany: Springer-Verlag, 2010.

[17] M. Drmota, W. Szpankowski, and K. Viswanathan, "Mutual information for a deletion channel," in *Proc. ISIT*, Jul. 2012, pp. 2561–2565.

[18] R. O'Donnell, *Analysis of Boolean Functions*. Cambridge, U.K.:Cambridge Univ. Press, 2014.

Yanjun Han received his B.Eng. degree with the highest honor in electronic engineering from Tsinghua University, Beijing, China in 2015. He is currently working towards the Ph.D. degree in the Department of Electrical Engineering at Stanford University, under the supervision of Prof. Tsachy Weissman. His research interests include high-dimensional statistics, nonparametric statistics, probability theory and information theory, with applications in communications, data compression, and learning.

Or Ordentlich received the B.Sc. degree (cum laude) in 2010, M.Sc. degree (summa cum laude) in 2011, and his PhD degree in 2016, all in electrical engineering in Tel Aviv University, Israel. He is currently a postdoctoral associate in the Laboratory for Information and Decision Systems at the Massachusetts Institute of Technology (MIT), Cambridge.

Or is the recipient of the MIT - Technion Postdoctoral Fellowship, the Adams Fellowship awarded by the Israel Academy of Sciences and Humanities, the Thalheimer Scholarship for graduate students, the Advanced Communication Center (ACC) Feder Family Award for outstanding research work in the field of communication technologies (2011,2014), and the Weinstein Prize for research in signal processing (2011, 2013, 2014).

Ofer Shayevitz received the B.Sc. degree (summa cum laude) from the Technion Institute of Technology, Haifa, Israel, in 1997 and the M.Sc. and Ph.D. degrees from the Tel-Aviv University, Tel Aviv, Israel, in 2004 and 2009, respectively, all in electrical engineering. He is currently a Senior Lecturer in the Department of EE - Systems at the Tel Aviv University, and also serves as the head of the Advanced Communication Center (ACC). Before joining the department, he was a postdoctoral fellow in the Information Theory and Applications (ITA) Center at the University of California, San Diego, from 2008 to 2011, and worked as a quantitative analyst with the D. E. Shaw group in New York from 2011 to 2013. Prior to his graduate studies, he served as an engineer and team leader in the Israeli Defense Forces from 1997 to 2003, and as an algorithms engineer at CellGuide from 2003 to 2004. Dr. Shayevitz is the recipient of the ITA postdoctoral fellowship (2009 - 2011), the Adams fellowship awarded by the Israel Academy of Sciences and Humanities (2006 - 2008), the Advanced Communication Center (ACC) Feder Family award (2009), and the Weinstein prize (2006 - 2009).